Helping Nomis find value at e-Car

May 2019

Team Beta - Stefano Zavagli, Jidapa Thanabhusest, Bosun Adebaki

MBA2965B - Data Science Applications in Finance and Accounting

Executive Summary

Overview: We are a team of consultants working with Nomis, a data analytics startup, to develop a plan to help e-Car, an internet based auto loan provider, transition to a data analytics based pricing strategy. Currently, only 22% of e-Car's loan applicants accept a loan offer from e-Car. This is below the industry standard. We attempt to employ different statistical methods to help the company maximize profits by 1) increasing the number of loan applicants that accept an auto loan offer once offered from the incredibly low 22%., 2) identifying the maximum loan interest rate that can be offered to customers who have accepted the auto loan offer.

Results:

We ran a K-Means Clustering Analysis to reduce the noise in the data and try to gain as much insight as possible from the statistics. We identified 9 key user clusters, i.e. user groupings, and used these separately to inform our analyses and recommendations. Furthermore, we trained 9 specific Decision Tree models for each of these clusters, and distilled further insight from the customer's decision making trends. Our goal was to focus on determining which users were higher prone to reject e-Car's loan offers, and address them with customized APR offerings. For the purpose of this written report, we focused our analysis on the 3 most significant clusters. The rest of the results are available on the complete Python notebook.

Conclusion and Recommendations:

After performing the customer segmentation exercise, we realized that some users decision making trends were difficult to discern, hence more careful experiments should be conducted in order to determine the proper APR threshold. Our final recommendations for e-Car rely on addressing the customer base on a cluster differentiated manner. We would recommend that e-Car segments customers along the 8 groups. For clusters 0 and 1, we would recommend that e-Car considers lowering the upper range of the APR it offers in order to entice more customers to accepting its loans. For cluster 5 and possibly cluster 0, we would recommend that e-Car considers increasing the APR offerings, as our analysis indicates that these groups have a higher willingness to pay. In addition to the aforementioned analysis, we would recommend that e-Car runs experiments on these and other clusters in order to better assess their willingness to pay.

Project Overview

Project goals

We are a team of consultants who are working with Nomis, a data analytics startup, and are developing a plan for e-Car's transition to a data-driven pricing strategy. The strategy should allow the company to maximize profits by 1) increasing the number of loan applicants that accept an auto loan offer once offered from the incredibly low 22%., 2) identifying the maximum loan interest rate that could be offered to customers (currently there is little price discrimination).

In doing so, we seek to identify potential inefficiencies in e-Car's current pricing strategies to provide Peter Thompson, the company's current CEO, with sufficient evidence to demonstrate that there's additional value to be captured by hiring Nomis to undertake more extensive analysis..

Company and industry background

The auto lending business in the US is incredibly competitive, representing between 15%-20% of total national consumer credit, with lenders providing financing to millions of americans each year. Customers can obtain auto loans via a) direct to bank / credit union process, b) indirect channels from auto dealers, c) internet.

e-Car is a specialized online auto lender that offers car loans to customers via the internet. It's customers arrive to e-Car's website by three main routes: 1) direct to the website, 2) indirectly from a partner website (e.g. AutoCompare or other used-car sites), 3) indirectly from other auto-loan comparison websites (e.g. lending tree). After obtaining information on an applicant's financial standing as well as details relating to the purpose of the specific loan request, the company either a) offers a loan that an individual has 45 days to accept, b) rejects the applicant as "too risky".

Whilst the company has been successful through historically offering loans at interest rates (annual percentage rates or APRs) that were generally unprofitable, the company has now gained sufficient product market fit to begin offering loans that are more competitive and profitable. Currently, approximately 22% of the loan offers it makes are accepted by applicants, with the loan *funded*. 78% of the loans offered to customers are not accepted e-Car treats this as *a lost sale*. Importantly, e-Car's existing pricing scheme has been created in such a way that *on any given day, any two customers in the same risk band, asking for a loan of the same type with the same term would typically be quoted the same APR.* As such, e-Car does not consider each customer's willingness to pay..

Given the company's historic approach of offering loans as unprofitable rates, e-Car is looking at how to segment customers such that it can better price optimize with respect to the APRs it offers on *funded* loans (i.e. understand the maximum price it can offer to these

customers before they reject). Additionally, such a segmentation would also help e-Car to reduce the percentage of *lost sales* through a more effective (cheaper) pricing strategy with respect to APRs (this based on the simplifying assumption that price is the only factor applicants choose to decide on).

Model development and operational details

From the data provided by e-Car, our team was able to construct a series of models that are able to explore and gain insight from the loan application process. Naturally, our models are data-driven, and as such, they use various statistical and mathematical methods to process and analyze the data. Before jumping into the data analysis, the team contemplated a variety of different model types, and ran a few experiments to evaluate the relevance of each particular model type.

First, we ran linear and logistic regressions; secondly, we ran decision trees; and finally, we performed experiments with clustering models. Particularly, we faced some obstacles with regards to the regression models, as it is especially difficult to model a discrete APR threshold for which customers will cross the line between accepting the loan offer vs. rejecting the loan offer. The features in the dataset did provide us with some statistical description of the consumer insights, but the results from the logistic and linear regressions were difficult to interpret, and ultimately they proved challenging to translate into a concrete business recommendation.

Hence, after completion of our model experiments, we decided to pursue a combination of clustering plus decision trees for the e-Car data analysis. This approach yielded the best results in terms of quality of insight, degree of complexity, and interpretability. As we will see in the next section, we avoided some of the most common problems, such as overfitting, by limiting the depth of the decision trees and the depth of our clustering model. A flow chart displaying a general overview of our model is reported in Exhibit 1. The most important metric for our model would be specificity since we care more about detecting people who want to reject or high true negative rate and then assume that we can change their decision by adjusting APR.

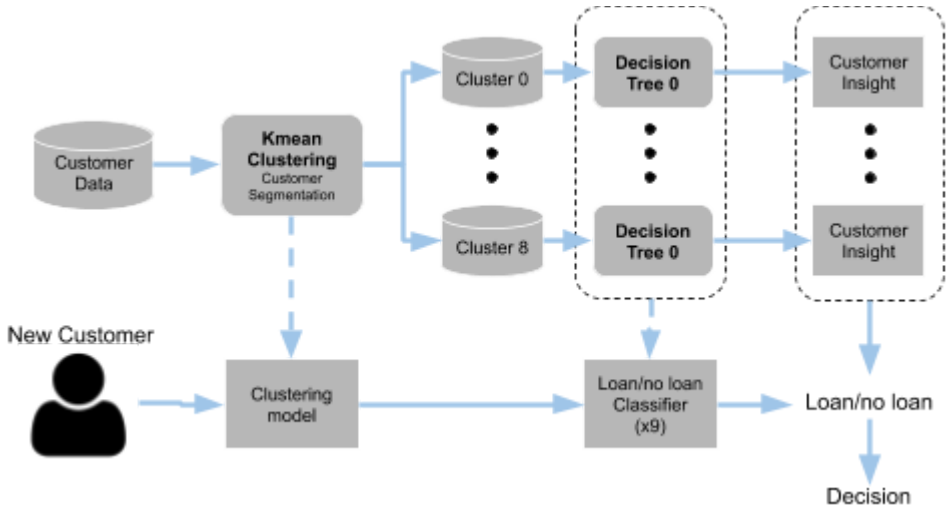

Exhibit 1: Complete model overview

Key Results

Overall description of the results

We start the results discussion by characterizing our dataset. The data gathered by e-Car was complete and clean, thus facilitating our work tremendously. The amount of information is substantial enough to guarantee good statistical representativeness; more precisely, the data included 208,805 records representing the outcome of all approved loan applications from January 2002 through December 2002.

That said, the data presented a modest challenge to our analysis, which was the unbalanced distribution of loan outcomes: only 45,787 users out of the total 208,805 accepted their loan offer, which is circa 22% of the total sample.

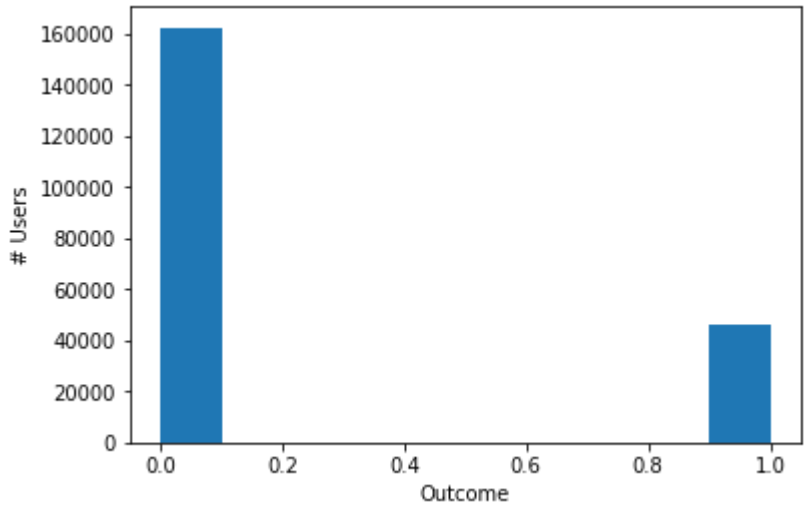

Exhibit 2: Bar chart of user outcomes (0 = reject offer, 1 = accept offer)

As mentioned in the model selection discussion, the result of the regression models did not give us a good amount of insight into the data, plus they were hard to interpret. Our idea behind using a clustering model was that of implementing regression analysis with the addition of performing a customer segmentation beforehand. This technique aids meaningfully in reducing a great deal of the noise in the data, thus improving our predictive capabilities. More specifically, we clustered our analysis in terms of focusing on specific groups of users/customers at a time.

The underlying method for execution the clustering technique was the K-means algorithm, and the steps of clustering construction are described as follows:

1. Convert categorical features into dummy binary variables
2. Drop features that are not directly customer's specific, such as Outcome, Rate, Competitor Rate, etc.
3. Scale the data to range between 0 and 1, as the clustering model requires minimizing the variation of the euclidean distance of the data among the same cluster. The data fields with larger values like Amount, could create unwanted biases in the model.
4. Perform a PCA to reduce the dimension of the features
5. Find the number of desired clusters by inspecting the plot of '*Error rate' vs. '# Clusters'* (as reported in Exhibit 3).

6. Clustering!
7. Visualize results

Following the sequence of steps described above, we were able to build a structure and that outputs a series of graphs and plots that aid us in understanding the dataset. The first thing we needed to do with our model was to choose the number of clusters that yielded the best results. This step is key in ensuring good quality from the analytical exercise. We selected the number of clusters by Elbow method or picking the point where increasing more clusters will decrease only a small amount of error in the 'Error rate' curve. This happens right before the number of clusters reaches 10. The concept of curve inflection choice is very similar to choosing the best tree depth for decision trees in cross validation.

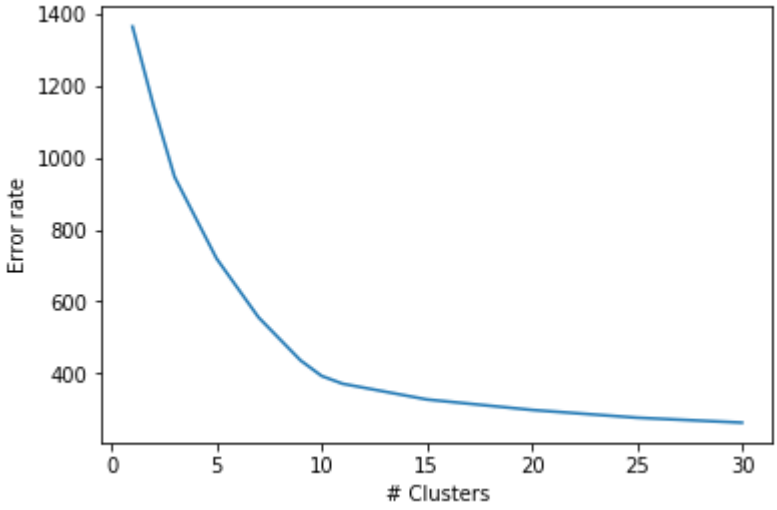

Exhibit 3: Graph for selection of the desired number of clusters

Once we constructed our model structure and debugged the program, we were finally able to analyze each of our clusters and draw insights from their data distributions. The final clustering arrangement looks like this:

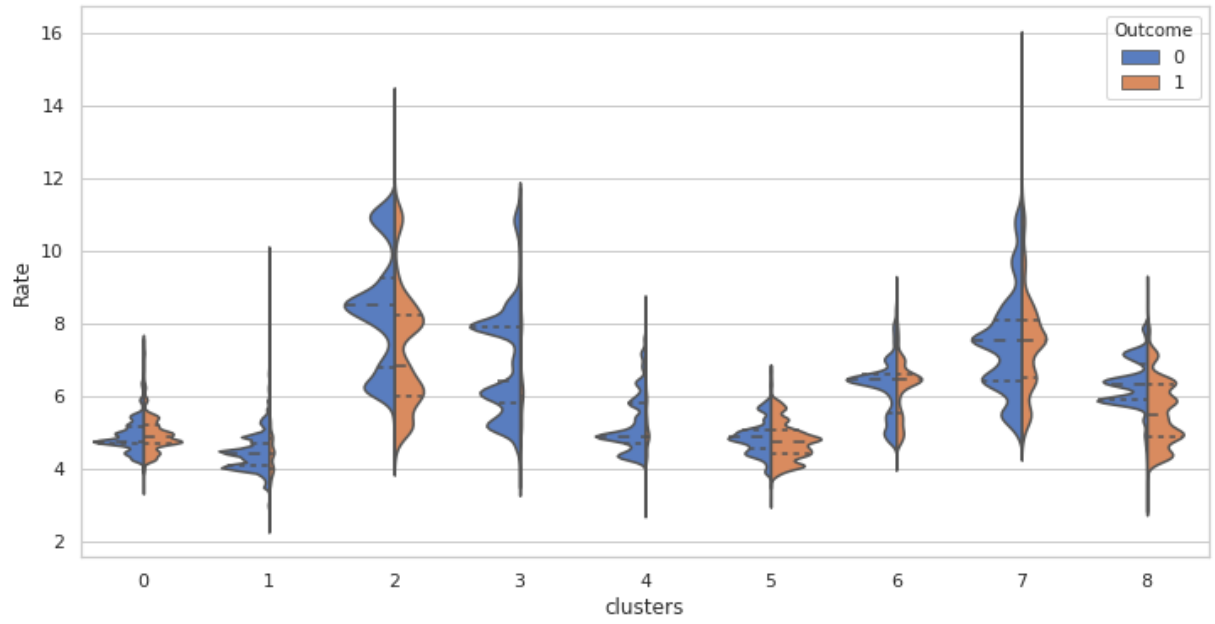

Exhibit 4: Distribution of APR 'Rate' according to customer cluster

(0 = reject, 1 = accept)

From Exhibit 4, we can already see very clearly some phenomena in the data. Clusters 1, 3 and 4 are heavily skewed towards rejecting the loan offers, while the rest of the clusters are more evenly balanced. We can infer from these distributions that e-Car is not capturing many customers in the categories belonging to clusters 1, 3 and 4. From this graph, we can also infer that there is a lot of room for improvement in the rest of clusters, that is if these user clusters effectively represent individuals that share a common set of traits that, and each one of them has the same propensity to participate in a loan contract. This assumption can seem slightly idealistic, yet on aggregate, our model will capture* statistical trends that can be interpreted at the individual level. Moreover, ideally we would like the majority of the cluster distributions to be colored in orange, as it would entail that e-Car is capturing the majority of customers in their respective classes.

We must be very mindful that these loan applications and users might not be fully committed to pursuing a loan contract. Some users might be comparing offers on a more informative note. Hence, it will not always be possible to capture a majority of the distribution as seen in the clusters - simply offering a lower APR to each customer class will not guarantee better revenue streams.

To better understand the composition of each cluster and the attributes that characterize each user class, we plotted the complete cluster-vs-feature set of distributions here below:

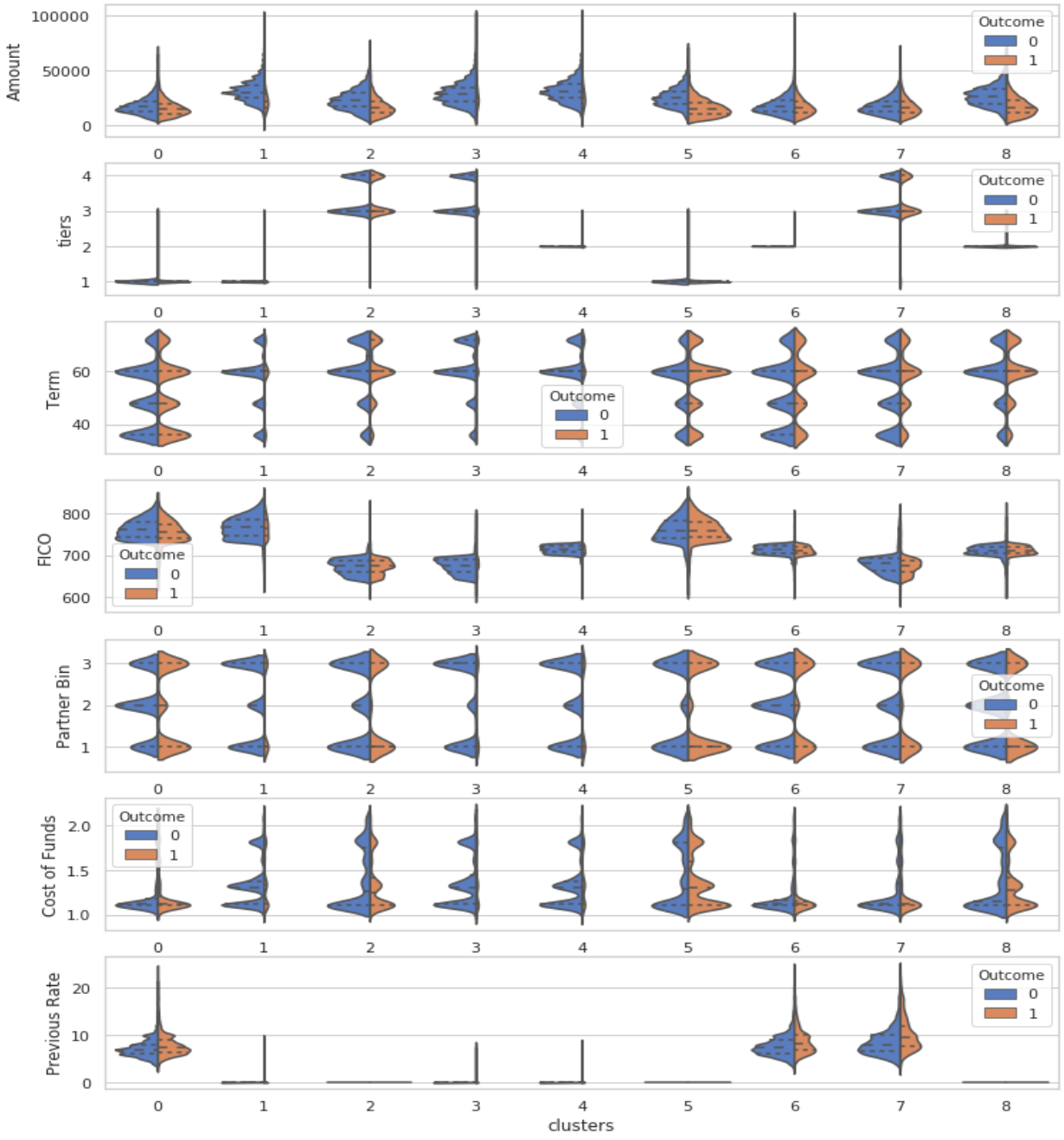

Exhibit 5: User distributions - 'cluster' vs 'features'

This exhibit shows the complete set of distributions that classified groups of users into specific clusters (numerated 0 through 8 on the x-axis). The vastness of the picture can

look overwhelming, but it is important that we illustrate the global results in one single exhibit so that the clusters can be fully understand. Moreover, we will examine the decompositions in depth as we delve into individual analyses of each cluster. Furthermore, hereinafter we coupled a specific decision tree with each cluster, as we attempt to identify the main driving factors that influence user decision making.

Starting chronologically, we will take a look at the most important clusters and provide our perspective and recommendations for each user class. We choose the 3 most significant clusters to present our detailed analysis of the data. The rest of the output results can be provided to you by contacting one of our representatives at the email address we will provide to you.

CLUSTER 0:

The user composition of this cluster can be explained by looking at the plots from Exhibit 5. From here on, we display a series of subplots derived from the global picture in Exhibit 5, and report them as new exhibits for the purposes of each cluster. These subplots are displayed as exemplified in Exhibit 6. The first subplot shows one of the differentiating features for users in this class: the loan amount for which these users are applying lies somewhat low relative to other clusters; 95% of the users in this cluster are asking for less than $38,000.

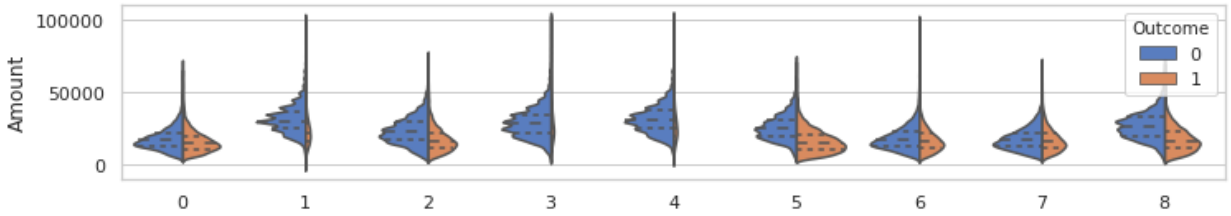

Exhibit 6: 'cluster' vs 'features' **- CLUSTER 0**

We can observe, primarily going back to Exhibit 5, but also from Exhibit 6, that the users in this cluster are generally applying for loan amounts in the lower range of the scale, have good risk band profile, high FICO score, and are planning to refinance a loan with respect to a previous rate. For this cluster of customers, e-Car is offering APRs concentrated between 4.1% and 6.3% (see Exhibit 4), resulting in a majority of users rejecting e-Car's offer. For this customer class, e-Car could consider lowering the upper range* of their APR offers, to try to engage more customers into their financing contracts.

Next, we examine the decision tree output for CLUSTER 0. We chose a decision tree depth according to the K-cross validation analysis, where the goal is to minimize the validation 'Error', while maintaining a degree of complexity as low as possible (to mitigate overfitting). A graph showing the deductive capabilities of our decision tree with respect to node depth is reported in Exhibit 7:

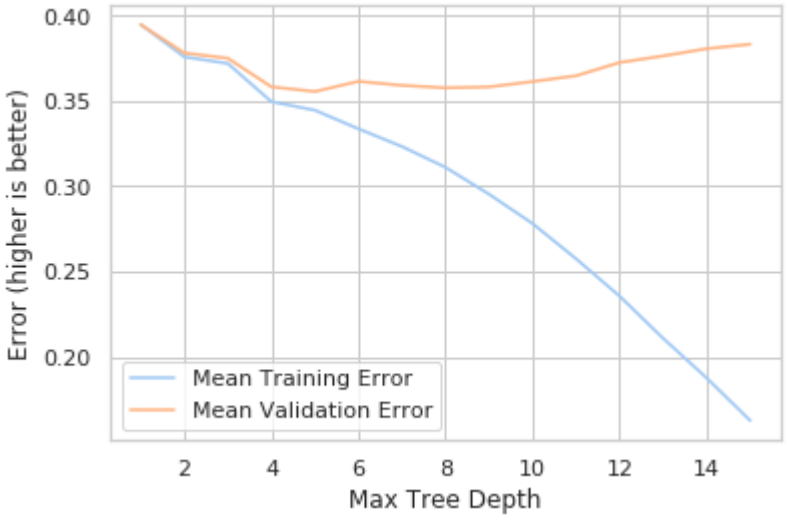

Exhibit 7: Selection of maximum depth for decision tree in **CLUSTER 0**

This method for "best depth analysis" was also repeated for the decision trees in each one of the other clusters (1-8). While all these "best depth analysis" are not presented in this report, the resulting decision trees are displayed as exhibits in the corresponding sections. The final decision tree and analogous confusion matrix for CLUSTER 0 are reported here:

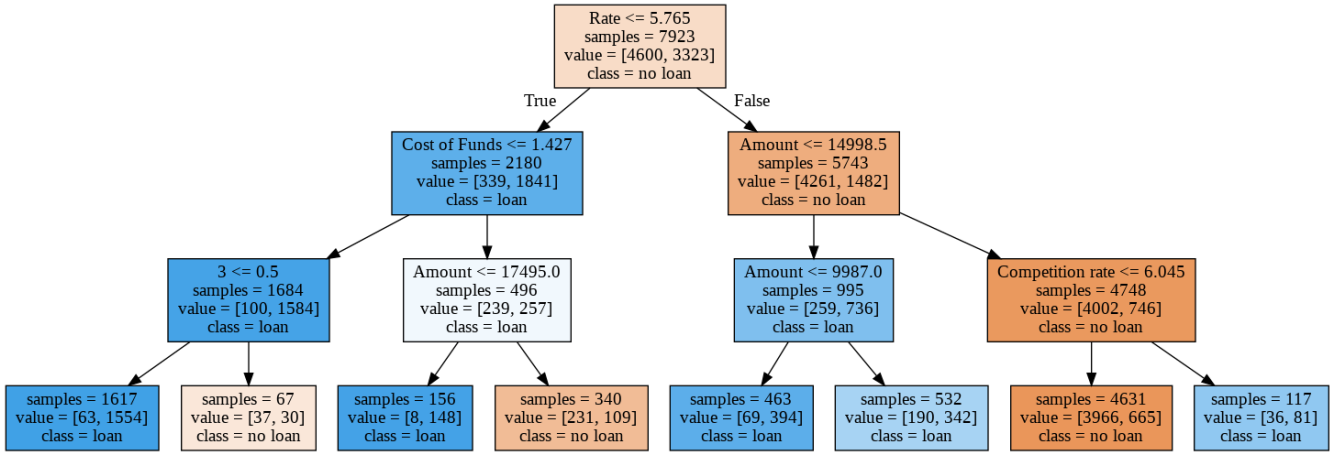

Exhibit 8: Decision tree for **CLUSTER 0**

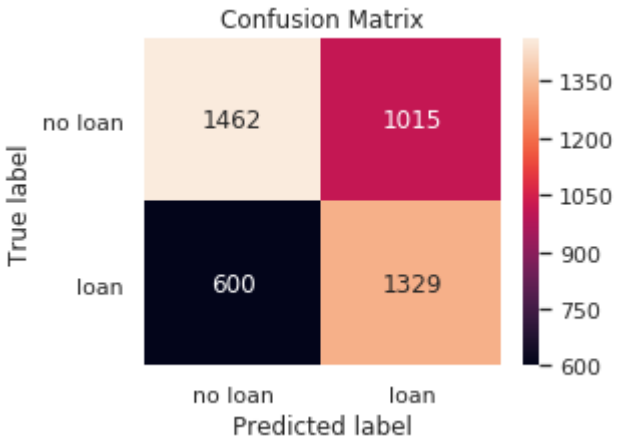

Exhibit 9: Confusion matrix for **CLUSTER 0**

From the confusion matrix, we can observe that the Sensitivity and Specificity metrics are not exceptionally deterministic. Sensitivity stands at 69%, while Specificity stands at 59%. Although the decision tree does not prove hugely predictive based on these two counts, we can examine some for the tree nodes to draw meaningful conclusions for e-Car.

We can immediately recognize the most important nodes based on their color intensity (see Exhibit 8). These represent branches and nodes that correspond to a significant differentiating factor. Specifically, 'Competition rate' and 'Tier' seem to be accounting for a large influence in the user's decision making. In cases where competitor's rate is lower than 6%, and e-Car offers 'Rate' <5.8% users tend to reject the offer. Our recommendation to e-Car is to keep in check the competitor's rate when offering loan contracts to potential customers in this class, and possibly adjust their APR offer accordingly.

CLUSTER 1:

Next we focus on cluster number 1. Once again, the user composition is illustrated in Exhibit 5, and we report Exhibit 10 to illustrate one of the primary differentiating features of this user class, the FICO score.

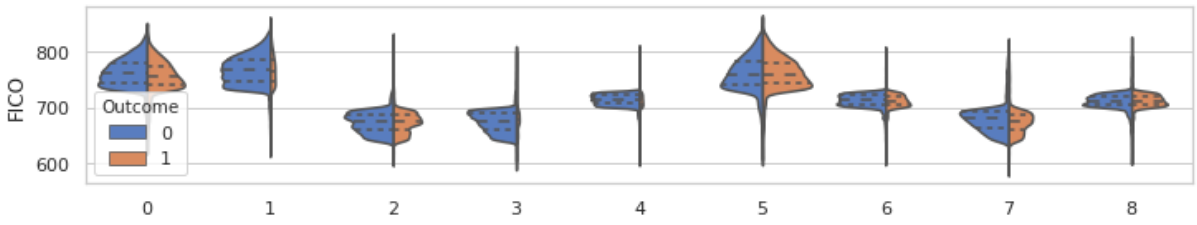

Exhibit 10: 'cluster' vs 'features' **- CLUSTER 1**

Users in this cluster have a relatively high FICO score, only 2 other clusters display a comparably high distribution. FICO values for this cluster typically range from 670 to 830, and these users are applying for loan amounts in the middle-to-high end of the range. These users have good risk band profile (Tier 1), and typically request term loans of 60 months. For this cluster, e-Car is offering APRs concentrated between 3.3% and 5.9% (see Exhibit 4), resulting in the vast majority of users rejecting e-Car's offer.

As with CLUSTER 0, for this customer class e-Car could consider lowering the upper range of their APR offers to try to appeal to more customers. While the APRs offered are already quite low, they are still above the 'Cost of Funds' (see Exhibit 11). We would strongly recommend e-Car to calculate the appropriate profit and margin scenarios before moving with the APR adjustments for this class..

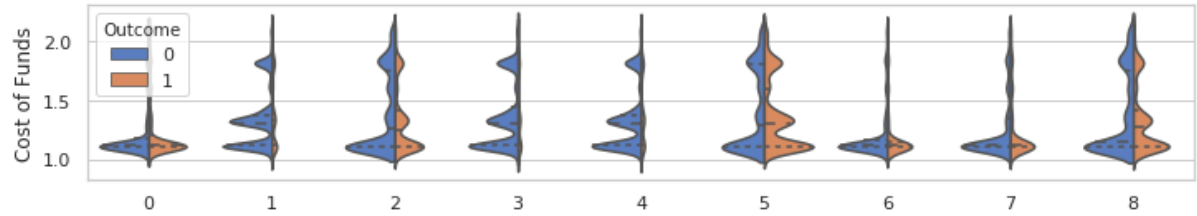

Exhibit 11: 'cluster' vs 'features' **- CLUSTER 1**

Additionally, the decision trees for this cluster show a large inclination in the decision making towards the 'Amount' factor. The nodes display a majority of rejection outcomes in correspondence to loan amounts north of ~$17000 with only 7% of users accepting e-Car's APR offer (see Exhibit 12). The Sensitivity and Specificity of this tree stand at 22% and 98% respectively, meaning that these driving features have good capability of predicting rejection outcomes (see Exhibit 13).

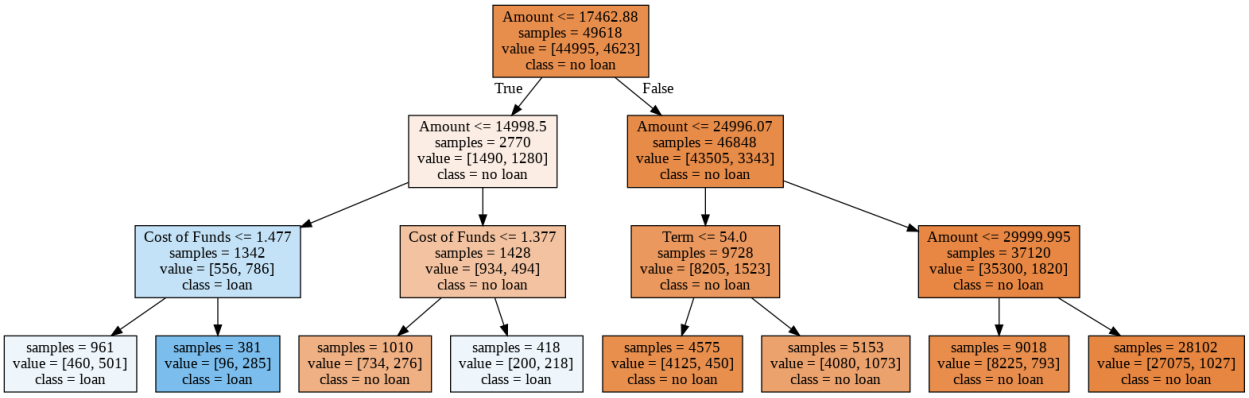

Exhibit 12: Decision tree for **CLUSTER 1**

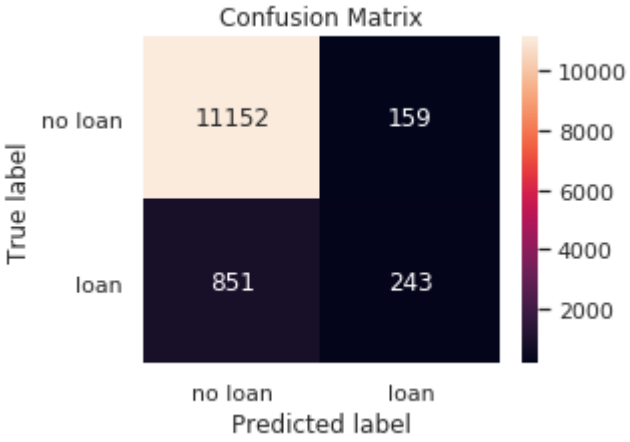

Exhibit 13: Confusion matrix for **CLUSTER 1**

CLUSTER 8:

For the last cluster, the interesting characteristic is that the acceptance rate is pretty high compared to the other clusters that have similar moderate FICO scores, for example CLUSTER 4 and CLUSTER 6 (see Exhibit 14). Further analysis of the features of CLUSTER 8 finds that another important characteristic is that the majority of customers came from 'Partner 1' and they had no 'Previous Rate', i.e. first time financing a purchase.

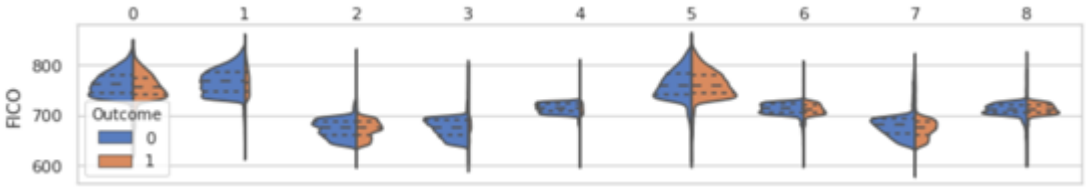

Exhibit 14: CLUSTER 1 - 'cluster' vs 'FICO'

More importantly, CLUSTER 8 shows a very visible APR cut-off threshold as shown in the decision tree in Exhibit 15. With the APR lower than a certain threshold, most users accept e-Car's offer. To analyze the threshold we fitted the data of users from CLUSTER 8 to a decision tree and the outcome is as follows:

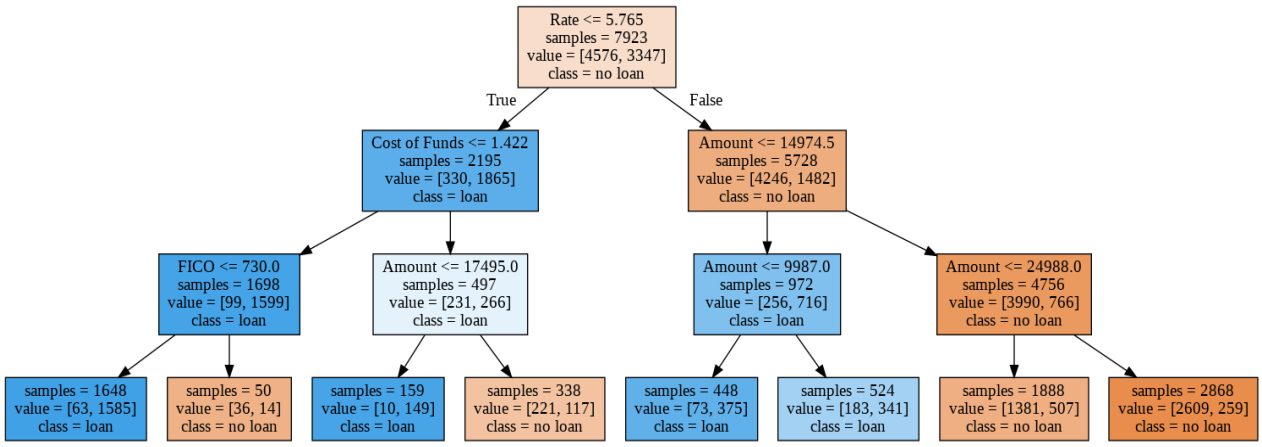

Exhibit 15: Decision tree for **CLUSTER 8**

The APR cutoff between users who rejected and accepted the loan offers can be clearly identified at the first node of the decision tree. We can see an even proportion with 4576 users out of 7923 rejecting e-Cars offer, vs. 3347 out of 7923 accepting the offer. From this result we can assume that e-Car would be able to convince some of the users that have a tendency to reject by simply decreasing the 'Rate' a few percentage points below 5.8%.

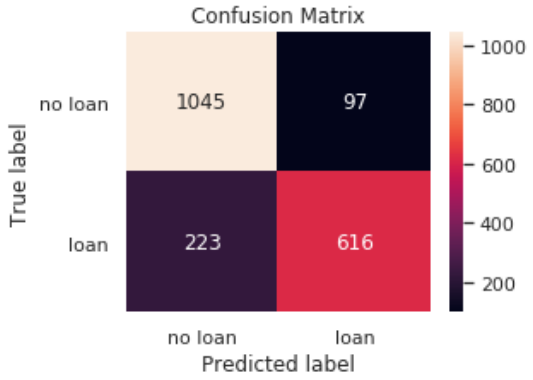

Exhibit 16: Confusion matrix for **CLUSTER 8**

On a parallel note, the Accuracy of CLUSTER 8's decision tree is roughly 85% and the Specificity is 91%. Users classified as True Negative, i.e. those who rejected our offer amount to a total of 1045; which supports our assumption regarding users that will change their mind when offered a lower APR, because at least 50% of them were predicted by the model.

Conclusions

Finally, the results our team found serve to highlight many of the key areas where e-Car is missing on potential customers. Our clustering model is particularly useful in segmenting the population into different groups with ranges of common features.

Conducting our analyzing on a class-by-class basis helps us identify the driving factors in the user's decision of accepting c.f. rejecting a loan offer. Furthermore, after our analysis we can make a series of recommendations to e-Car that will help the company capture a higher proportion of potential customers that are utilizing e-Car's online platform:

- e-Car can utilize the clustering model provided by Nomis Solutions to better address their client base and seek to improve their loan customization process.
- We notice a substantial room for improvement in terms of existing customer segments that have not been fully exploited. High proportions of the individual cluster distributions are not accepting the current loan offerings by e-Car.
- A couple of clusters (5 and 0) show margin for higher APR pricing experiments. This would enable e-Car to increase profit on the same loans, while ensuring to keep most of the customers. This resilience trait is key in order to extract higher profits from the existing customer base.

- That said, most of the remaining user clusters are in need of lower APR offers in order for e-Car to capture a greater proportion of the addressable market
- e-Car can now utilize our user classes and quantify the sensitivity of different groups to ensure that their marketing and investment efforts are only directed to customers that really have upside potential.
- Any pricing strategy must still be analyzed from a cost of capital and operating expenses perspective before being implemented globally.

Future Recommendations

Given the data, resources and time constraints, the analysis by Nomis Solutions seeked to achieve a qualitative conclusion with regards to the dataset. While at the moment we are not able to pinpoint the exact APR offering that each individual user would accept, our team believes that the insights from this report will aid e-Car in focusing their efforts on the right subsets of customers, thus reducing costs and increasing top line revenue. Furthermore, we believe that given more data and time, we can assist e-Car in optimizing their pricing strategies so as to maximize company profits. Ultimately, the models we utilized in this exercise had the objective of displaying a qualitative yet actionable result. A future collaboration between e-Car and Nomis Solutions could develop more exact numerical recommendations. Some of the approaches that Nomis could pursue are: fitting the data with Random Forests or implementing Gradient Boosting. Rebalancing of the dataset is also and effective yet simple technique that could further distill our results. While using other types of models entails a trade off between interpretability and performance, the higher complexity models would allow us to analyze the cut-off for APR and strike a value that optimizes for e-Car's profits.

For the full Python notebook, please check out the link below.

https://colab.research.google.com/drive/1cyNpAokJ5fc14vj5HADdkjkh0eIRYQbt

Learning Component

Core learnings during the project

We learned that it can often be difficult to find a model that fits the available data whilst also meeting the broad project objectives. To solve this, it's important for the data scientists and the business people to think clearly about the broad range of insights required.. For example, we had initially considered using some simple regressions to generate broad insight, however, the results were relatively meaningless.

In order to create competitive pricing strategies, clustering is a great initial tool that can be used to segment customers into groups that have moderately similar characteristics, both from a pure data science perspective and a human readability perspective.

What team challenges did you overcome?

Aside from the fairly common struggle to coordinate calendars for a mixed data/business team that has a broad range of commitments outside of this class, one of the biggest challenges we had to overcome was getting right the balance between the technical detail and the broader business implications. For example, identifying a number of statistically significant clusters is useful, however, it can prove not material if we're unable to translate this into meaningful business decisions.

Further, because the business focus was primarily on the macro implications, we struggled at the beginning to create a clear plan of where to start and were a bit overwhelmed by the number of data analytics techniques that could be used.

Company Dynamics: What industry/company problems did you overcome? Future recommendations: What would you do differently in the future?

In the future, we would want to look and analyze data both on the customers that e-Car accepted and those that it rejected providing loans to. This additional data would help us analyze whether the company really is prioritizing the correct features.

Given that in practice, APR is only just one of the drivers for a customer's decision to accept an e-Car loan, we would also want to do some testing to identify what non-APR adjustments to the application process could increase the proportion of clients who accept loans, the ability to increase the APRs.